[Ve]tting [R]esponse [I]ntegrity from cross-[T]alk dependencies in [A]dversarial [S]urveys: *"Who can catch a liar?"*

Robert Gibbons, Ishanu Chattopadhyay[1,4,5,7]★

[1]Department of Medicine, University of Chicago, Chicago, IL 60637, USA
[4]Committee on Quantitative Methods in Social, Behavioral, and Health Sciences, University of Chicago, Chicago, IL 60637, USA
[5]Committee on Genetics, Genomics & Systems Biology, University of Chicago, Chicago, IL 60637, USA

★To whom correspondence should be addressed: e-mail: ishanu@uchicago.edu.

Abstract: **Malingering[1,2] or faking the symptoms of a mental health disorder can confound structured diagnostic interviews and hinder clinical psychiatric assessments. We introduce an artificial intelligence (AI) framework for detecting symptom fabrication in Post-Traumatic Stress Disorder (PTSD) diagnoses, for which malingering is a known problem[3,4], partially ascribable to the potential for secondary financial gain from positive diagnoses. Algorithm `VeRITAS` employs novel generative AI to infer statistical dependencies inherent in true response patterns, and flags responses which violate these subtle constraints. Mimicing these emergent patterns is difficult even with psychiatric training, resulting in a robust mechanism for recognizing fabrications. With a study sample of $n = 624$ patients, `VeRITAS` is estimated to have a Area Under the Curve (AUC) of $\geqq 95.0\%$, with one sensitivity/specificity pair of 92% and 95% respectively. Our tool offers an objective, disease-specific, fast (average time $\leqq 2$ min) approach to detect fake PTSD, and if adopted, can ensure that healthcare resources and disability concessions reach those genuinely in need, while helping to maintain integrity of clinical data. Moreover, the ability to identify and help patients who might be malingering due to other mental health conditions, poverty or socio-economic compulsions can improve general health outcomes in disadvantaged communities.**

INTRODUCTION

Diagnosis of mental health disorders typically rely on structured interviews[5] susceptible to intentional fabrication of symptoms[1,2], often referred to as "malingering". In the specific context of clinically diagnosing Post-Traumatic Stress Disorder (PTSD), subjective stressors, a high degree of similarity in presentation, the realtively easy access to information on how to fake PTSD, and financial incentives from a positive diagnosis in legal procedures or disability claims, is known to hinder accurate assessments. Here, we present an approach based on a generative model to flag malingering suspects rapidly, efficiently and accurately; essentially demonstrating a generative artificial intelligence (AI) framework that, paraphrasing Ekman[6], "*can catch a liar*".

Clinically, PTSD is an anxiety disorder that can develop after experiencing a traumatic event. In the United States, substantial disability compensation may be available for those with mental disorders, which while being an crucial resource for the truely afflicted, incentivizes malingering[3,4]. Thus, faking PTSD to access medical treatment, commit insurance, personal injury and other frauds, or in an attempt to evade criminal liability and penalties[7–10] is unfortunately not rare. While PTSD is a serious mental health condition associated with substance use disorder, mood disorder, anxiety disorder, personality disorder, increased morbidity, and possibly with increased mortality[11,12], false diagnoses can cause substantial financial drain[13,14] to healthcare systems, divert crucial resources from where they are needed[15], and interfere with study outcomes by introducing inaccuracies in clinical data[16]. Accurate disambiguation of true and fake PTSD is therefore of high importance, especially with sources suggesting that over 20% of personal injury cases, as well as 20% of the Veterans seeking combat compensation could be fabricating their condition[3,17–19]. In addition to curbing financial fraud, ability to flag such incongruities can help address the myriad of factors that can drive malingering behavior, including other mental health conditions, a lack of access to healthcare[20,21] and an inability to seek help arising from poverty and other broad-ranging socio-economic conditions.

Despite the general difficulty in formulating principles to detect malingering[22], multiple standardized tests[23,24] and validity assessment tools[25] have been proposed, with limited success. These tools typically aim to incorporate

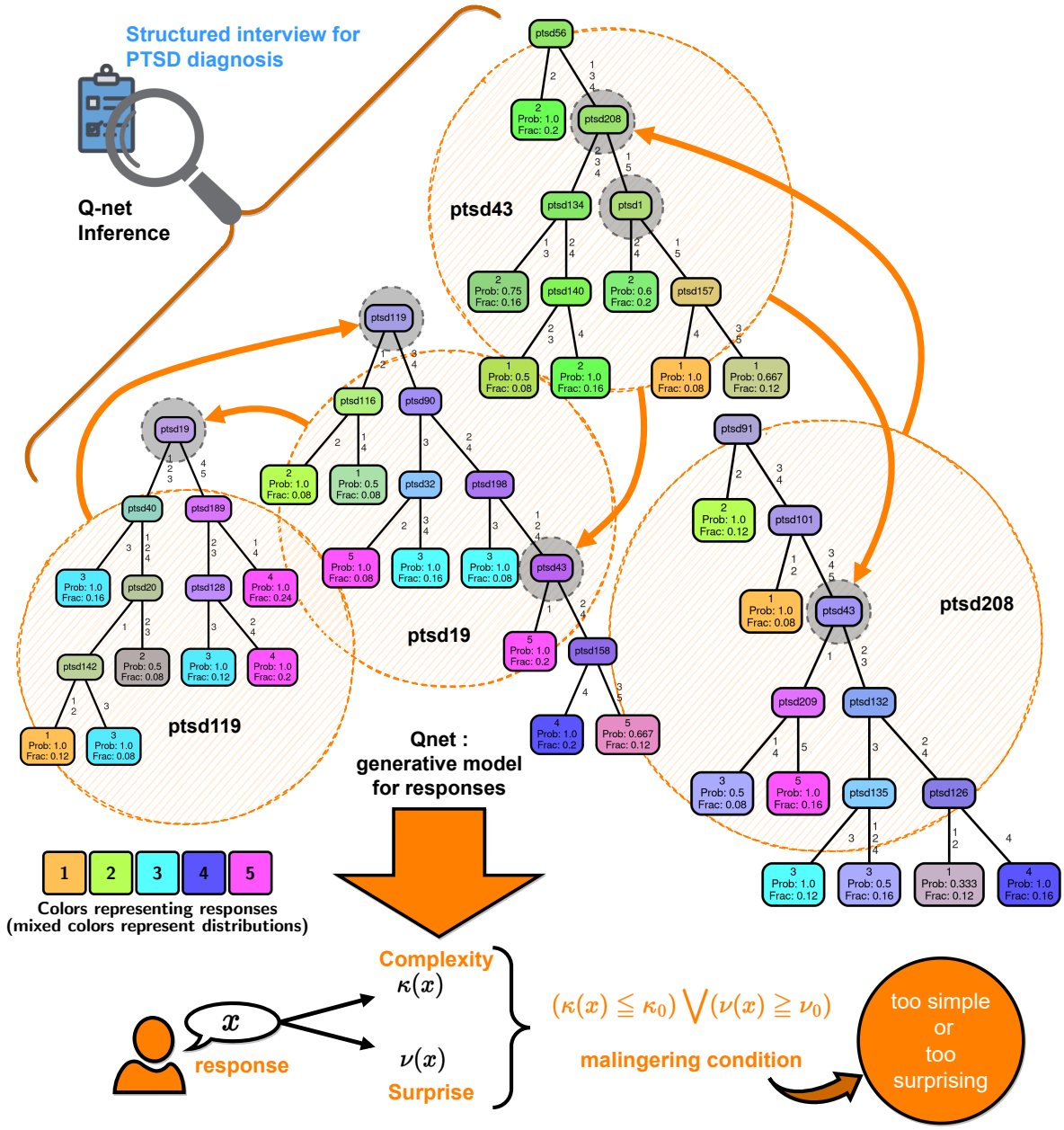

Fig. 1: **Conceptual framework.** Using a dataset of responses to a validated structured interview for PTSD diagnoses, along with physician-validated clinical diagnoses, we infer generative models for responses for PTSD patients. In our framework for detecting malingering, we flag responses as those which are highly "surprising" (defined as violating infered cross-dependencies between individual response items) or are too simple (lacking complexity in the response patterns typical of non-malingered responses). The precise "malingering condition" shown above is validated from theoretical considerations as well as field experimental data.

patterns observed in diagnostic populations that might disambiguate faked symptoms from real ones, or ask similar or related questions multiple times to verify consistency. However, existing approaches do not target specific disorders, almost always require expert interpretation, are subjective, and by design are unlikely to be effective against a malingerer with psychiatric training (See Table 2). Other strategies with physiological monitoring and linguistic analysis[26–29] cannot be easily adopted in structured or semi-structured interviews.

In VeRITAS we leverage the fact that responses in a structured interview have statistical dependencies arising from the nature of the questions themselves, and are modulated by the trait we are aiming to detect *e.g.*, PTSD pathology. We operationalize this principle without requiring human-understanding of the specific items being presented to the subject; thus making the approach specific to the disease at hand (PTSD), while being potentially generalizable to other disorders if appropriate training data is available.

Our key finding here is that the subtle cross-dependencies between the interview items are challenging to mimic on-the-fly, even if the subject is knowledgeable about how such patients tend to respond *i.e.* with training in the mental health services. Thus, VeRITAS offers a robust approach to verifying response validity, can target

TABLE 1: Prolific Dataset Participant Characteristics and Success Rate as a percentage of the representation, at 90% sensitivity

characteristics	Time taken	Age	Count	success rate (%)
Race: Asian	188.4	34.3	24	8.3
Race: Black	329.5	38.0	33	9.1
Race: Mixed	195.8	31.4	20	5.0
Race: Other	286.2	38.5	13	0.0
Race: White	186.5	42.7	220	1.8
Sex: Female	201.5	40.5	167	2.4
Sex: Male	213.5	41.0	141	4.3
Residence: United Kingdom	213.5	43.1	110	2.7
Residence: United States	202.9	39.4	200	3.5

specific disorders, can be administered in less that 4 minutes, and require little subjective interpretation.

RESULTS

Participants and Data Sources

Our first dataset (referred to as the VA dataset) comprises $n = 304$ participants recruited at a Veterans Health Administration facility for an earlier study[30]. Veterans between the ages of 18 and 89 years were recruited with written informed consent. Once eligibility was determined by the study team, participants completed a PTSD-symptom questionnaire from the CAT-PTSD item bank[30,31], comprising 211 items, including some items from the PTSD Checklist (PCL-5). Participants were also interviewed using the Clinician-Administered Scale for PTSD for Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), which resulted in 86 PTSD diagnoses, and 218 participants deemed as not having PTSD. We used 60% of the data for inferring our generative models, and the rest was used for validation, including determining the out-of-sample AUC for identifying PTSD vs no-PTSD cases. The possibility of malingering was not recorded in this dataset.

The second dataset generated in the course of the present study is from online surveys conducted by a third party vendor (Prolific) hired by the study team. Referred to here as the Prolific dataset, it comprises $n = 310$ participants (200 in the US, and 110 in the UK), screened for the absence of past or present mental health diagnoses. The participants were asked to fake symptoms of PTSD while taking the VeRITAS interview.

In total, this study considered $n = 624$ participants comprising US Veterans, and drawn from the general population in the US and the UK, with broad representation across sexes and ethnicities. Average completion times and detailed demographic composition of the respondents is shown in Table 1.

Item Presentation and Response Sequence in Administration

In VeRITAS we present a small random subset of the full item bank to the subjects. Although the models for cross-dependency are inferred using the complete response sequence to all 211 items available, presenting a smaller item set makes the test feasible, while also producing a vast number of possible variations. In the version we validated, 20 items are presented. These are randomly generated from a distribution reflecting the impact of the items in the inferred models on their ability to disambiguate between responses from diagnosed PTSD patients and the control group. Impact is measured using standard SHAP analysis of the inferred models (See Methods for details). SHAP-values for items are normalized to a probability mass function (pmf), which is then sampled to produce the item set. We can select items adaptively similar to CAT-PTSD[30]; such complications have been presently avoided.

Integrated Diagnostic Capability

Given a sequence of responses to a diagnostic interview, our first task is to determine if a particular subject should be diagnosed with PTSD, if the possibility of malingering is ignored. We call this a "naive diagnosis". This diagnostic information might be available to VeRITAS externally (*e.g.* from a physician's assessment of the patients). However, VeRITAS also has an integrated capability for naive diagnosis: we identify separate generative models for 1) the diagnosed set of patients in the training set (M^+, some of them might be malingering) and 2) for the patients identified as not having PTSD (M^0). Then given the sequence of responses from a new subject, we estimate if that sequence is more likely to have been generated by the model for M^+ vs that for M^0. We validate this diagnostic capability using the training data described above, and we achieve good disambiguation between M^+ and M^0, with out-of-sample AUC 0.867 ± 0.008, which is at par or better compared

to reported tools, *e.g.*, CAT-PTSD[32]. In the VeRITAS algorithm, for a given response sequence $\mathbf{x}$, the naive diagnosis risk score is denoted as $\mu(x)$, and referred to as simply the "score" (See Methods for definition and computational approach).

Principle of Characterizing Malingered Responses

Amongst those who have a naive PTSD diagnosis, we aim to flag subjects with a quantifiable high estimated likelihood of malingering. Note that the in training set we only have the naive diagnoses information; thus some participants diagnosed with PTSD in the dataset could have been malingering, but are not flagged to be so. To flag malingering sunjects, our first insight is that such response sequences have high average "surprise", *i.e.* deviate more on average from the context-specific model predictions of item responses. Our other insight is that malingering tend to generate less complex patterns in the response sequences. We understand complexity in the formal sense: more complex objects are less compressible, and random sequences are not significantly compressible. Thus a sequence of all 1's is very compressible and therefore non-random, but a sequence generated from teh sequential tosses of a fair coin is not very compressible, and is highly random. We hypothesize that true response sequences will tend to have maximal randomness, conditional on being constrained by the emergent cross-dependencies. Thus, true responses should have the just right degree of identifiable structure, and no more. More structure will make it less random, and less structure will increase surprise, and thus too much and too little structure are both indicative of malingering. Both of these criteria are based on statistical properties of the response sequence, and do not require a human understanding of the items. Quantitatively, for each response sequence $\mathbf{x}$, we compute two quantities: $\kappa(x)$ and $\nu(x)$, referred to as the complexity and the surprise parameters respectively.

Thus, κ, ν, μ are random variables which are functions of the response sequence of the subject, and have distributions that may be characterized from the VA data, despite not having designation of "malingering" in that dataset. This is because that these quantities are computable from just the response sequence themselves. Concretely, we propose that a subject with a response sequence $\mathbf{x}$ is malingering if for a suitably chosen thresholds μ_0, ν_0, κ_0:

$$\big(\mu(x) \geqq \mu_0\big) \bigwedge \Big(\big(\kappa(x) \leqq \kappa_0\big) \vee \big(\nu(x) \geqq \nu_0\big) \Big) \tag{1}$$

This may be paraphrased as that a response sequence has high likelihood of malingering if it 1) produces a naive diagnosis of PTSD with high probability, and 2) is either too surprising or too simple. The decision thresholds are obtained from theoretical considerations and the VA data.

VeRITAS Validation Strategy

Since the VA data does not not direct indication of either presence of absence of malingering, we adopted an non-standard approach for validating VeRITAS. We assumed that the Prolific participants do not have PTSD, and that all of them were attempting to malinger. This allows us to directly measure false negative rates, as function of the VeRITAS parameters. Then we check how many of the VA participants are flagged as malingering amongst the ones with a naive PTSD diagnosis. To determine a lower bound on performance we can assume all of these subjects are false positives (which is unlikely, but nevertheless is an upper bound on false positives as function of VeRITAS parameters). This allows us to construct a lower envelop of the ROC curve, and hence estimate a lower bound of the AUC.

Using this approach, we obtain a minimum AUC of 95%.

Q-net Inference and Cross-talk Modeling

Training and Validation

Performance

Comparison Against State of Art

DISCUSSION

Malingering remains a persistent problem, especially in psychiatric and criminal justice settings, with prevalence estimates ranging from $8-64\%$[33–35]. Here we develop a novel system to address this gap for the need for further development of reliable, rigorous, and principled methods for detecting deception remains highly significant[36].

In this study, we introduce a novel algorithmic technique for identifying deception in structured interviews. This approach first infers a detailed model of the response patterns of previous respondents, then uses a measure of consistency known as dissonance to identify if future respondents' answers are indicative of deception.

Our model-based, data-driven approach is notably distinct from traditional methods of malingering detection, which are typically constructed for specific contexts[37]. in the context of psychiatric diagnosis detection of

faked symptoms is often based on standardized tests; some commonly used ones are summarized in Table 2. Tests such as the Structures Interview of Reported Symptoms (SIRS)[23], Structured Inventory of Malingered Symptomology (SIMS)[24], and the Minnesota Multiphasic Personality Inventory-2 and its associated validity scales[25] draw extensively on research of both legitimate psychiatric patients and malingerers, and employ a combination of detection strategies focused on symptoms that are either unlikely to be presented by genuine patients, or that tend to be amplified in malingerers[38]. For detection of cognitive malingering, physiological, arousal-based approaches such as the Control Question Technique[39], and cognitive load-inducing approaches such as requiring subjects to perform a concurrent task ($e.g.$ maintaining eye contact) and the Autobiographical Implicit Association Test (aIAT) have been suggested.

Accuracy rates for these research-driven state of the art detection techniques are typically in the range of 85-95% depending on the specific problem context[40–42] (similarly high sensitivities have also been reported[24]). However, among the limitations to these methods is the requirement of substantial domain research and/or subject matter expertise to develop. Thus while these methods perform well in the specific context for which they were developed, it may be challenging and/or expensive to broadly extend to detection in other contexts. In the case of cognitive malingering methods, vulnerabilities to clinician/adminstrator bias, high rates of false positives and false negatives, and vulnerability to coaching have been noted as possible limitations.

In contrast with these methods, our results demonstrate that it is possible to achieve similar (or perhaps improved) performance in detecting malingering by identifying structural differences in an individual's responses compared to a set of baseline responses. The basic insight underpinning our results is that consistent, genuine responses to interview questions are typically not independent. In general, there are dependence patterns between interview questions, which are often nontrivial and difficult or impossible to identify a priori. Given the entirety of responses an individual has provided in an ongoing interview, the response probabilities for a consistent response to the next question are governed by these dependence patterns. As these probabilistic constraints exist between all questions in the interview, a complex interaction network is induced over the high-dimensional response space. To infer the structure of this network, we use a novel architecture - the Q-net, a nonparametric generative model of the response network which reflects the n-way inter-question relations found amongst previous interview respondents.

From the model's approximation of the probabilistic dependencies present in the response network, we derive the dissonance as an intuitive, individually computable measure of response consistency. For all data sets in this study, mean dissonance values for random malingering and expert malingering are significantly higher than for actual responses, empirically demonstrating its utility for identifying such responses. As should be expected, we also found that the expert malingering responses were on average less dissonant than random malingering, showing that domain knowledge does allow individuals to respond in a more convincing manner. However, their failure to adequately account for detailed inter-question dependencies tends to make their responses significantly different from actual responses in a measurable, quantitative sense.

The classification results obtained further demonstrate the ability of this framework to identify subtle deviations in the response patterns of actual and malingering respondents. Due to the lower dissonance generated by expert malingerers, we found it more difficult to identify such responses; however, in practice it appears unlikely that even the most knowledgeable domain experts would have the requisite quantitative knowledge of response distributions of the target phenotype to be able to respond in the manner of our idealized expert malingerers. Thus in real-world usage where true malingerers are likely a mixture of these two idealizations, we would expect performance to be bounded above and below by the empirical performance obtained for the random and expert scenarios.

In general, data-driven methods for detection of malingering are not as ubiquitous as in other domains; however, the increasing availability of modern sensory data collection has allowed for implementation of machine learning algorithms and computer vision techniques[43**?**,44] for detecting deception based on patterns in language, facial expressions, body movements, and eye movements[45,46]. These methods have been shown to outperform traditional lie detection methods in some cases, obtaining accuracy rates of over 90% for lie detection from patterns in facial expressions and body movements[6,45] and 60-80% using analysis of gaze direction and pupil dilation[46]. However, in many instances it is not feasible to collect detailed sensory data, possibly limiting the general applicability of these kinds of methods. By exclusively utilizing response data, our dissonance-based approach is not subject to such limitations.

In contexts such as these, where malingerers seek to obtain a target diagnosis, training data from past respondents often includes a clinical phenotype for each respondent. However, this is not always available (as was observed for the CCHHS data set). Nonetheless, in the absence of the underlying diagnostic phenotypes, our approach still appears to perform well under a scenario in which expert malingerers are also unsure of ground-truth diagnoses.

In other types of structured interviews, such as internet surveys and opinion polls, built-in data validity checks are frequently used to attempt to proactively identify disingenuous respondents. Among such checks, unusually fast response times have been observed to be indicative of poor data quality[47,48], demonstrating a lack of attention

by interview respondents[49]. Our finding of a negative correlation between dissonance and response time adds some support to this notion. However, it has been noted that duration-based validity checks may be vulnerable to a high false negative rate[50], suggesting a need for most robust measures of assessing validity.

METHODS

1. DEFINITIONS & NOTATION

Definition 1 (Survey)**.** *A survey for the purpose of this work is a structured interview, consisting of a finite number of questions (items) posed to a set of participants, with these items drawn from a finite item bank, and whose responses must be one froma pre-specified set fo choices, e.g., the Likert scale, with missing values for the responses allowed.*

Definition 2 (Response vector)**.** *A response vector is the set of responses to a survey from a single participant, typically assuming that not all items are posed, and allows for the possibility that some responses are missing.*

A Q-net, as described here, is a model of the response dependency structure for questions (items) posed to participants in a survey. The Q-net explicitly estimates individual conditional distributions of each item response, which collectively serve as a model of the full joint distribution of the responses.

Definition 3 (Q-net)**.** *Let $X \sim P$ be an n-dimensional discrete random vector supported on a finite set Σ and following distribution P, i.e.*

$$X = (X_1, \ldots, X_n) \sim P, \qquad \mathrm{supp}(X) = \Sigma = \prod_{i=1}^{n} \Sigma_i \quad \textit{with } |\Sigma| < \infty.$$

For $i = 1, \ldots, n$, let $P_i := P(X_i \mid X_j = x_j \text{ for } j \neq i)$ denote the conditional distribution of X_i given the values of the other components of X. Finally, for each $i = 1, \ldots, n$, let Φ_i^P denote an estimate of the distribution P_i. Then the set $\Phi^P := \{\Phi_i^P\}_{i=1}^n$ is called a Quasinet (Q-net)*. Identifying the true distribution P as the one describing the joint statistics of the responses from a survey with n items, we also refer to Φ^P as the Q-net for the survey P.*

When P is clear from context, we may omit the superscript and simply write $\Phi = \{\Phi_i\}$ to denote the Q-net. The motivation for Definition 3 is that the collection of all estimators $\Phi = \{\Phi_i\}$ contained in a Q-net represents the set of all inferred dependencies from the observed ecosystem. While the definition allows for arbitrary method of algorithm to construct the estimators Φ_i, the utility of a Q-net clearly depends primarily on the properties of the Φ_i. In this study, we aim to minimize the set of a priori assumptions on the overall model structure to allow the complex dependencies present in P to emerge. To that end, throughout this work all Q-nets are computed using conditional inference trees[51] (a variant of classification and regression trees) to compute each Φ_i. In general each Q-net component Φ_i is computed independently from the other Φ_j, which allows a network structure to emerge amongst these estimators.

An important quantity for an inferred Q-net is the persistence function $\omega_{\mathbf{x}}$.

Definition 4 (Persistence Function)**.** *Given a survey P inducing the Q-net Φ^P and a response vector $\mathbf{x} = (x_1, \ldots, x_n)$, the persistence $\omega_{\mathbf{x}}$ of $\mathbf{x}$ in the population modeled by the Q-net:*

$$\omega_{\mathbf{x}}^P := \mathrm{Pr}(\mathbf{x} \in P) = \prod_{i=1}^{n} \Phi_i^P(X_i = x_i \mid X_j = x_j, j \neq i) \tag{2}$$

The persistence function $\omega_{\mathbf{x}}^P$, as the name suggests, is the probability that $\mathbf{x}$ persists, *i.e.*, $Pr(\mathbf{x} \to \mathbf{x})$ for the population modeled by the Q-net P, with $1 - \omega_{\mathbf{x}}^P$ being the probability that $\mathbf{x}$ is altered by a random perturbation.

We will show that if for two inferred Q-net models P, Q, we have $\omega_{\mathbf{x}}^P \geqq \omega_{\mathbf{x}}^Q$, then it is more likely that model P generated $\mathbf{x}$. This is an important result that justifies the definition of the score parameter in Defn. 6.

The Q-net allows us to rigorously compute bounds on the probability of a spontaneous change from one response vector to another, induced by spontaneous chance variations (Fig **??**). Not all perturbations in a vector are either likely or contextually meaningful. With an exponentially exploding number of possibilities in which a vector over a large set of items can vary, it is computationally intractable to directly model all possible dependecies; nevertheless, we can constrain the possibilities using the patterns we uncover via the Q-net construction. A key piece of this approach is to design an intrinsic distance between response vectors, which is reflective of this underlying dependency structure.

Definition 5 (q-distance)**.** *Let $\Phi^P = \{\Phi_i^P\}_{i=1}^n$ and $\Phi^Q = \{\Phi_i^Q\}_{i=1}^n$ denote Q-nets on populations P and Q, and suppose $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ are samples of $X \sim P$ and $Y \sim Q$ respectively. Then the*

q-distance $\theta_{P,Q}(\mathbf{x},\mathbf{y})$ between $\mathbf{x}$ and $\mathbf{y}$ is

$$\theta_{P,Q}(\mathbf{x},\mathbf{y}) := \frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{J}^{\frac{1}{2}}\left(\Phi_i^P(X_i|X_j=x_j, j\neq i)\,\|\,\Phi_i^Q(Y_i|Y_j=y_j, j\neq i\right)\right]$$

where $\mathbb{J}$ denotes the Jensen-Shannon divergence[52].

For brevity, we may write simply θ (dropping the suffixes) if the populations are clear from context. Since the Jensen-Shannon distance $\mathbb{J}$ is a legitimate metric[53] on the set of probability distributions (unlike KL-divergence), θ inherits nonnegativity, symmetry, and respects the triangle inequality; it follows that q-distance is a (pseudo)-metric on Σ. Note that, being a pseudo-metric implies that we may have $\theta(\mathbf{x},\mathbf{y})=0$ for $\mathbf{x}\neq\mathbf{y}$, i.e. distinct vectors can induce the same distributions over each index, and thus have zero distance. This is in fact desirable, since we do not want our distance to be sensitive to changes that are not meaningful. The intuition is that not all variations are equally important or likely. Moreover, we show in Theorem 1 that the log-likelihood of a vector $\mathbf{x}$ transitioning to $\mathbf{y}$ scales with $\theta(\mathbf{x},\mathbf{y})$, allowing us to directly estimate the probability of spontaneous (or sequential) jumps between abundance profiles.

Theorem 1 (Probability Bound)**.** *Given a vector $\mathbf{x}$ of length n from P that transitions to $\mathbf{y}$ from Q, we have the following bounds at significance level α.*

$$\omega_y e^{\frac{\sqrt{8}N^2}{1-\alpha}\theta(\mathbf{x},\mathbf{y})} \geqq Pr(\mathbf{x}\to\mathbf{y}) \geqq \omega_{\mathbf{y}} e^{-\frac{\sqrt{8}N^2}{1-\alpha}\theta(\mathbf{x},\mathbf{y})} \tag{3}$$

where $\omega_{\mathbf{y}}$ is the persistence of $\mathbf{y}$ (Def. 4), and $\theta(\mathbf{x},\mathbf{y})$ is the q-distance between $\mathbf{x},\mathbf{y}$ (Def. 5).

Proof. See later in Section 3. □

Theorem 1 gives theoretical backing to the claim that samples generated by the Q-net indeed reflect likely perturbation possibilities from the current state. Thus we can use the Q-net to draw contextually realistic samples that respect the cross dependencies and reduce surprise (that is, the Q-net-inferred conditional distributions can be used to generate approximate samples from the population P). This has several implications, such as the ability to easily handle missing/incomplete data.

Remark 1 (Neighborhood Structure)**.** *It follows from Th. 1 that we have for some constant C,*

$$\ln\left|\frac{Pr(x\to y)}{Pr(y\to y)}\right| \leqq C\theta(x,y) \tag{4}$$

implying for all response vectors y within a small neighborhood of x (small in metric θ), we have:

$$Pr(y\to x) \approx Pr(x\to x) \tag{5}$$

which reveals an important special structure on local neighborhoods.

2. VeRITAS Analysis

Definition 6 (Algorithm `VeRITAS` Parameters)**.** *We introduce three parameters referred tro as teh complexity, surprise and score parameters (κ,ν,μ respectively) for a given response vector x:*

$$\textit{complexity:}\quad \kappa \triangleq -\frac{1}{|x|}\ln Pr(x\to x|M^+) = -\frac{\ln\omega_x^{M^+}}{|x|} \tag{6}$$

$$\textit{surprise:}\quad \nu \triangleq \mathbf{E}_i\left(1-\Phi_i^{M^+}(x_{-i})|_{x_i}\right) \tag{7}$$

$$\textit{score:}\quad \mu \triangleq \frac{\ln Pr(x\to x|M^+)}{\ln Pr(x\to x|M^0)} = \frac{\ln\omega_x^{M^+}}{\ln\omega_x^{M^0}} \tag{8}$$

where M^+ indicates the sub-population exhibiting a particular trait of interest e.g. a mental health disorder such as PTSD, and M^0 is the control sub-population where this trait is absent.

Interpretation of the Parameters

We

Definition 7 (Malingering property)**.** *A response vector x is defined to have the malingering property if:*

$$\left(\mu(x)\geqq\mu_0\right)\bigwedge\left(\left(\kappa(x)\leqq\kappa_0\right)\vee\left(\nu(x)\geqq\nu_0\right)\right) \tag{9}$$

Set of malingering responses is denoted as $\mathscr{M}$.

The decision thresholds κ_0,ν_0,μ_0 are inferred from survey data.

Lemma 1 (Complexity)**.** *For a survey with n items, and assuming L to the number of possible responses to each item, the unconditional probability of a response vector x occurring among all feasible responses is bounded above by $(e^{\kappa}/L)^n$, where $\kappa(x)$ is the complexity parameter for response $\mathbf{x}$.*

Proof. Let $\kappa(x) \leqq \kappa'$. From Def. 6, we have for a response vector x,

$$-\frac{1}{n}\ln \omega_x \leqq \kappa' \Rightarrow \omega_x \geqq e^{-n\kappa'} \tag{10}$$

Summing on both sides over all responses x with $\kappa(x) \leqq \kappa'$ (assume there are N_x such sequences), we have:

$$1 \geqq \sum_x \omega_x \geqq \sum_x e^{-n\kappa'} \tag{11}$$

where the first inequality follows from observing that responses very close to x in the q-distance metric have a specific structure, namely $\omega_x \approx Pr(y \to x)$ (See Remark 1) and responses further away have smaller jump probabilities, which then implies:

$$N_x \sum_x e^{-n\kappa'} \leqq 1 \Rightarrow N_x \leqq e^{n\kappa'} \tag{12}$$

The result then follows from noting that the complete set of possible responses has the size L^n. □

Lemma 1 justifies why a low value of κ implies the possibility of an un-natural response, because the odds of generating such a response is remarkably small.

Lemma 2 (Surprise)**.** *For any response vector x, we have:*

$$\nu(x) \leqq 1 - e^{-\kappa(x)} \tag{13}$$

Proof. Denoting $\Phi_i(x_{-i})|_{x_i}$ as a_i, we note that $\omega_x^{1/n}$ is the geometric mean of the vector of a_is, while $\mathbf{E}_i\left(\Phi_i(x_{-i})|_{x_i}\right)$ is the arithmetic mean of the same vector, which then completes the proof by noting:

$$-\mathbf{E}_i\left(\Phi_i(x_{-i})|_{x_i}\right) \leqq -\omega_x^{1/n} \Rightarrow \nu(x) \leqq 1 - \omega_x^{1/n} \tag{14}$$

□

Why the Defined Property Identifies Malingering

Lemma 2 indicates that the requirement of an upper bound on the surprise and a lower bound on the complexity are both aiming to flag responses which are unlikely to appear when the data (responses) are being generated by the underlying process corresponding to the phenotype of interest (PTSD). When such unlikely responses do appear appear nevertheless, it is likely that they are not being generated by the correct underlying process. Can it correspond to another "natural" process, *e.g.* a different mental disorder that is not intentional fabrication? If the score is higher than

Note that the remaining condition $\mu(x) \geqq \mu_0$ is a diagnosis criterion for the trait of interest (M^+), and may be replaced with a different condition if available for identifying participants with the M^+ trait. This particular form follows from a straightforward Bayesian argument on estimating the posterior.

3. Proof of Theorem 1

Theorem 2 (Probability bound)**.** *Given a sequence x of length N that transitions to a strain $y \in Q$, we have the following bounds at significance level α.*

$$\omega_y^Q e^{\frac{\sqrt{8}N^2}{1-\alpha}\theta(x,y)} \geqq Pr(x \to y) \geqq \omega_y^Q e^{-\frac{\sqrt{8}N^2}{1-\alpha}\theta(x,y)} \tag{15}$$

where ω_y^Q is the membership probability of strain y in the target population Q (See Def. **??***), and $\theta(x,y)$ is the q-distance between x, y (See Def. 5).*

Proof. Using Sanov's theorem[52] on large deviations, we conclude that the probability of spontaneous jump from strain $x \in P$ to strain $y \in Q$, with the possibility $P \neq Q$, is given by:

$$Pr(x \to y) = \prod_{i=1}^{N}\left(\Phi_i^P(x_{-i})|_{y_i}\right) \tag{16}$$

Writing the factors on the right hand side as:

$$\Phi_i^P(x_{-i})|_{y_i} = \Phi_i^Q(y_{-i})|_{y_i}\left(\frac{\Phi_i^P(x_{-i})|_{y_i}}{\Phi_i^Q(y_{-i})|_{y_i}}\right) \tag{17}$$

we note that $\Phi_i^P(x_{-i})$, $\Phi_i^Q(y_{-i})$ are distributions on the same index i, and hence:

$$|\Phi_i^P(x_{-i})_{y_i} - \Phi_i^Q(y_{-i})_{y_i}| \leqq \sum_{y_i \in \Sigma_i} |\Phi_i^P(x_{-i})_{y_i} - \Phi_i^Q(y_{-i})_{y_i}| \tag{18}$$

Using a standard refinement of Pinsker's inequality[54], and the relationship of Jensen-Shannon divergence with total variation, we get:

$$\theta_i \geqq \frac{1}{8}|\Phi_i^P(x_{-i})_{y_i} - \Phi_i^Q(y_{-i})_{y_i}|^2 \Rightarrow \left|1 - \frac{\Phi_i^Q(y_{-i})_{y_i}}{\Phi_i^P(x_{-i})_{y_i}}\right| \leqq \frac{1}{a_0}\sqrt{8\theta_i} \tag{19}$$

where a_0 is the smallest non-zero probability value of generating the entry at any index. We will see that this parameter is related to statistical significance of our bounds. First, we can formulate a lower bound as follows:

$$\log\left(\prod_{i=1}^{N}\frac{\Phi_i^P(x_{-i})|_{y_i}}{\Phi_i^Q(y_{-i})|_{y_i}}\right) = \sum_i \log\left(\frac{\Phi_i^P(x_{-i})|_{y_i}}{\Phi_i^Q(y_{-i})|_{y_i}}\right) \geqq \sum_i \left(1 - \frac{\Phi_i^Q(y_{-i})_{y_i}}{\Phi_i^P(x_{-i})_{y_i}}\right) \geqq \frac{\sqrt{8}}{a_0}\sum_i \theta_i^{1/2} = -\frac{\sqrt{8}N}{a_0}\theta \tag{20}$$

Similarly, the upper bound may be derived as:

$$\log\left(\prod_{i=1}^{N}\frac{\Phi_i^P(x_{-i})|_{y_i}}{\Phi_i^Q(y_{-i})|_{y_i}}\right) = \sum_i \log\left(\frac{\Phi_i^P(x_{-i})|_{y_i}}{\Phi_i^Q(y_{-i})|_{y_i}}\right) \leqq \sum_i \left(\frac{\Phi_i^Q(y_{-i})_{y_i}}{\Phi_i^P(x_{-i})_{y_i}} - 1\right) \leqq \frac{\sqrt{8}N}{a_0}\theta \tag{21}$$

Combining Eqs. 20 and 21, we conclude:

$$\omega_y^Q e^{\frac{\sqrt{8}N}{a_0}\theta} \geqq Pr(x \to y) \geqq \omega_y^Q e^{-\frac{\sqrt{8}N}{a_0}\theta} \tag{22}$$

Now, interpreting a_0 as the probability of generating an unlikely event below our desired threshold ($i.e.$ a "failure"), we note that the probability of generating at least one such event is given by $1-(1-a_0)^N$. Hence if α is the pre-specified significance level, we have for $N >> 1$:

$$a_0 \approx (1-\alpha)/N \tag{23}$$

Hence, we conclude, that at significance level $\geqq \alpha$, we have the bounds:

$$\omega_y^Q e^{\frac{\sqrt{8}N^2}{1-\alpha}\theta} \geqq Pr(x \to y) \geqq \omega_y^Q e^{-\frac{\sqrt{8}N^2}{1-\alpha}\theta} \tag{24}$$

□

TABLE 2: Summary of malingering/deception detection methods.

Approach/Tool	**Description**	**Noted Accuracy/Performance**
Structures Interview of Reported Symptoms (SIRS)[23]	An interview-based measure with multiple detection strategies.	No specific accuracy rate mentioned, but noted as a robust instrument.
Structured Inventory of Malingered Symptomology (SIMS)[24]	A paper-and-pencil screening device for detecting malingering with a sensitivity for detecting malingering of 95.6% in a study with college students.	95.6% sensitivity in a specific study context.
Minnesota Multiphasic Personality Inventory-2 (MMPI-2)[25]	A self-report measure assessing personality and psychopathology. Certain validity scales were developed to uncover malingering.	Not specified, but noted flaws and potential for false positives.
Millon Clinical Multiaxial Inventory MCMI-III	A self-report scale focusing on personality disorders.	No specific accuracy rate mentioned.
Miller Forensic Assessment of Symptoms (M-FAST)	A brief screening measure for malingered mental illness in forensic settings.	Noted issues with low internal consistency on some scales.
Human Lie Detectors	Not a specific tool, but the general method of humans attempting to discern lies from truth.	People are generally poor lie detectors.
Arousal-Based Approaches (like the Polygraph)	Techniques that rely on physiological responses.	Criticized for poor validity and high rate of false positives.

Cognitive Load-Inducing Approaches	Techniques that view deception as a cognitive act that generally imposes greater cognitive load on respondents than honesty does.	No specific accuracy rate mentioned.
Autobiographical Implicit Association Test (aIAT)[40]	Designed to determine whether respondents possess actual autobiographical memories.	91% accuracy rate in identifying genuine autobiographical memories.
Timed Antagonistic Response Alethiometer (TARA)[41]	A computer-administered, response time-based method of lie detection.	85% accuracy rate.
Detecting Faked Identities With Unexpected Questions and Mouse Movements	Technique using computer mouse movements in conjunction with unexpected questions to uncover faked identities[42]	95% accuracy rate.
Time-Restricted Integrity Confirmation (TRI-Con)	A cognitive load-inducing technique with potential to uncover different kinds of deception including malingering.	Up to 89% accuracy rate.
Activation-Decision-Construction-Action Theory (ADCAT)	A theory of high-stakes deception.	No specific accuracy rate mentioned, but this is more of a theoretical foundation rather than a specific tool or method.

REFERENCES

[1] Rogers, R. An overview of malingering and its assessment. *Psychiatric Clinics of North America* **20**, 15–27 (1997).

[2] Rogers, R. E. *Clinical assessment of malingering and deception* (Guilford Press, 2008).

[3] Frueh, B., Grubaugh, A., Elhai, J. & Buckley, T. Us department of veterans affairs disability policies for posttraumatic stress disorder: administrative trends and implications for treatment, rehabilitation, and research. *Am J Public Health* **97**, 2143–5 (2007).

[4] Taylor, S., Frueh, B. C. & Asmundson, G. J. G. Detection and management of malingering in people presenting for treatment of posttraumatic stress disorder: Methods, obstacles, and recommendations. *Journal of Anxiety Disorders* **21**, 22–41 (2007).

[5] Ali, S., Jabeen, S. & Alam, F. Multimodal approach to identifying malingered posttraumatic stress disorder: A review. *Innovations in Clinical Neuroscience* **12**, 12 (2015).

[6] Ekman, P. & O'Sullivan, M. Who can catch a liar? *American Psychologist* **46**, 913–920 (1991).

[7] Guriel, J. & Fremouw, W. Assessing malingered posttraumatic stress disorder: A critical review. *Clinical Psychology Review* **23**, 881–904 (2003).

[8] Salloway, S., Southwick, S. & Sadowsky, M. Opiate withdrawal presenting as posttraumatic stress disorder. *Hospital and Community Psychiatry* **41**, 666–667 (1990).

[9] Resnick, P. J., West, S. & Payne, J. W. Malingering of posttraumatic disorders. In Rogers, R. (ed.) *Clinical assessment of malingering and deception*, 109–127 (Guilford Press, 2008), 3 edn.

[10] Burkett, B. G. & Whitley, G. *Stolen valor: How the Vietnam generation was robbed of its heroes and history* (Verity Press, 1998).

[11] Goldstein, R. B. *et al.* The epidemiology of dsm-5 posttraumatic stress disorder in the united states: results from the national epidemiologic survey on alcohol and related conditions-iii. *Social psychiatry and psychiatric epidemiology* **51**, 1137–1148 (2016).

[12] Schnurr, P. P., Lunney, C. A., Bovin, M. J. & Marx, B. P. Posttraumatic stress disorder and quality of life: Extension of findings to veterans of the wars in iraq and afghanistan. *Clinical psychology review* **29**, 727–735 (2009).

[13] LoPiccolo, C., Goodkin, K. & Baldewicz, T. Current issues in the diagnosis and management of malingering. *Ann Med* **31**, 166–174 (1999).

[14] Oboler, S. Disability evaluations under the department of veterans affairs. In Rondinelli, R. & Katz, R. (eds.) *Impairment Rating and Disability Evaluation*, 187–217 (W. B. Saunders, Philadelphia, PA, 2000).

[15] Taylor, S. *Clinician's Guide to Treating PTSD: A Cognitive-Behavioral Approach* (Guilford Press, New York, 2006).

[16] Rosen, G. Dsm's cautionary guideline to rule out malingering can protect the ptsd data base. *J Anxiety Disorders* **20**, 530–535 (2006).

[17] Marx, B. & Holowka, D. Ptsd disability assessment. *PTSD Res Q* **22**, 1–6 (2011).

[18] Rogers, R., Sewell, K. & Goldstein, A. Explanatory models of malingering: a prototypical analysis. *Law & Hum Behav* **18**, 543–52 (1994).
[19] Lees-Haley, P. Mmpi-2 base rates for 492 personal injury plaintiffs: implications and challenges for forensic assessment. *J Clin Psychol* **53**, 745–55 (1997).
[20] Park, L., Costello, S., Li, J., Lee, R. & Jacobson, K. C. Race, health, and socioeconomic disparities associated with malingering in psychiatric patients at an urban emergency department. *General Hospital Psychiatry* **71**, 121–127 (2021).
[21] Muntaner, C., Eaton, W. W., Miech, R. & O'campo, P. Socioeconomic position and major mental disorders. *Epidemiologic reviews* **26**, 53–62 (2004).
[22] Drob, S. L., Meehan, K. B. & Waxman, S. E. Clinical and conceptual problems in the attribution of malingering in forensic evaluations. *The journal of the American Academy of Psychiatry and the Law* **37**, 98–106 (2009).
[23] Wong, S. & O'Sullivan, M. The structured interview of reported symptoms (sirs): An overview. *Assessment* **12**, 289–307 (2005).
[24] Smith, G. P. & Burger, G. K. Detection of malingering: validation of the structured inventory of malingered symptomatology (sims). *Journal of the American Academy of Psychiatry and the Law Online* **25**, 183–189 (1997).
[25] Ben-Porath, Y. S. *Interpreting the mmpi-2-rf* (U of Minnesota Press, 2012).
[26] Ekman, P. & O'Sullivan, M. Who can catch a liar? *American Psychologist* **46**, 913 (1991).
[27] Mihalcea, R. & Strapparava, C. The lie detector: Explorations in the automatic recognition of deceptive language. In *ACL-IJCNLP 2009* (2009).
[28] Burgoon, J. K., Blair, J. P. & Strom, R. E. Cognitive biases and nonverbal cue availability in detecting deception. *Human Communication Research* **34**, 572–599 (2008).
[29] Zhou, L., Burgoon, J. K., Nunamaker Jr, J. F. & Twitchell, D. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group Decision and Negotiation* **13**, 81–106 (2004).
[30] Brenner, L. A. *et al.* Development and validation of computerized adaptive assessment tools for the measurement of posttraumatic stress disorder among us military veterans. *JAMA Network Open* **4**, e2115707–e2115707 (2021).
[31] Gibbons, R. D. *et al.* Development of a computerized adaptive test for depression. *Archives of general psychiatry* **69**, 1104–1112 (2012).
[32] Brenner, L. A. *et al.* Development and validation of computerized adaptive assessment tools for the measurement of posttraumatic stress disorder among us military veterans. *JAMA Network Open* **4**, e2115707 (2021). URL https://doi.org/10.1001/jamanetworkopen.2021.15707.
[33] McDermott, B., Dualan, I. & Scott, C. Malingering in the correctional system: does incentive affect prevalence? *Int'l J L & Psychiatry* **36**, 287–92 (2013).
[34] Schmidt, T., Krüger, M. & Ullmann, U. Base rate of probable malingering and its indicators in the assessment of mental disorders-retrospective analysis of a sample of forensic psychological evaluations. *Die Rehabilitation* **59**, 231–236 (2020).
[35] Matto, M., McNiel, D. E. & Binder, R. L. A systematic approach to the detection of false ptsd. *The journal of the American Academy of Psychiatry and the Law* **47**, 325–334 (2019).
[36] DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L. & Charlton, K. Cues to deception. *Psychological Bulletin* **129**, 74–118 (2003).
[37] Walczyk, J. J., Sewell, N. & DiBenedetto, M. B. A review of approaches to detecting malingering in forensic contexts and promising cognitive load-inducing lie detection techniques. *Frontiers in psychiatry* **9**, 700 (2018).
[38] Rogers, R. & Correa, A. A. Determinations of malingering: Evolution from case-based methods to detection strategies. *Psychiatry, Psychology and Law* **15**, 213–223 (2008).
[39] Vrij, A. & Mann, S. Who killed my relative? police officers' ability to detect real-life high-stake lies. *Psychology, Crime & Law* **7**, 119–132 (2001).
[40] Sartori, G., Agosta, S., Zogmaister, C., Ferrara, S. D. & Castiello, U. How to accurately detect autobiographical events. *Psychological science* **19**, 772–780 (2008).
[41] Gregg, A. P. When vying reveals lying: The timed antagonistic response alethiometer. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* **21**, 621–647 (2007).
[42] Monaro, M., Gamberini, L. & Sartori, G. The detection of faked identity using unexpected questions and mouse dynamics. *PloS one* **12**, e0177851 (2017).
[43] Chan, E. Y., Danker-Hopfe, H. & Riemann, D. A machine learning approach to polysomnographic sleep study for lie detection. *IEEE Journal of Biomedical and Health Informatics* **22**, 716–722 (2018).
[44] Wang, X., Yang, Y., Zhang, J., Wang, L. & Liu, S. A machine learning approach to lie detection based on eye movement and gesture analysis. *IEEE Transactions on Human-Machine Systems* **50**, 266–273 (2020).
[45] Yu, H., Chen, Y. & Fu, X. Deep learning for deception detection: a comprehensive review. *IEEE Transactions on Information Forensics and Security* **13**, 2615–2629 (2018).

[46] Ducharme, J. J., Langleben, D. D. & Verma, R. Eye tracking in the detection of deception: a review of the literature. *Psychology, Crime & Law* **23**, 875–895 (2017).
[47] Malhotra, N. Completion time and response order effects in web surveys. *Public opinion quarterly* **72**, 914–934 (2008).
[48] Callegaro, M., Yang, Y., Bhola, D. S., Dillman, D. A. & Chin, T.-Y. Response latency as an indicator of optimizing in online questionnaires. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* **103**, 5–25 (2009).
[49] Greszki, R., Meyer, M. & Schoen, H. Exploring the effects of removing "too fast" responses and respondents from web surveys. *Public Opinion Quarterly* **79**, 471–503 (2015).
[50] Kennedy, C. *et al.* Assessing the risks to online polls from bogus respondents. *Pew Research Center* **18** (2020).
[51] Sarda-Espinosa, A., Subbiah, S. & Bartz-Beielstein, T. Conditional inference trees for knowledge extraction from motor health condition data. *Engineering Applications of Artificial Intelligence* **62**, 26–37 (2017).
[52] Cover, T. M. & Thomas, J. A. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)* (Wiley-Interscience, New York, NY, USA, 2006).
[53] Fuglede, B. & Topsoe, F. Jensen-shannon divergence and hilbert space embedding. In *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, 31 (IEEE, 2004).
[54] Fedotov, A. A., Harremoës, P. & Topsoe, F. Refinements of pinsker's inequality. *IEEE Transactions on Information Theory* **49**, 1491–1498 (2003).